Towards Lifelong Learning Agents

Mikel Malagón[1], Jon Vadillo[1], Josu Ceberio[1], and Jose A. Lozano[1,2]

[1] University of the Basque Country (UPV/EHU), Donostia-San Sebastian, Spain
[2] Basque Center for Applied Mathematics (BCAM), Bilbao, Spain
{mikel.malagon,jon.vadillo,josu.ceberio,ja.lozano}@ehu.eus

Abstract. Over the past decade, Reinforcement Learning (RL) has led to significant breakthroughs in the field of Artificial Intelligence (AI), transitioning from academic benchmarks to real-world deployment. However, standard RL agents experience catastrophic performance degradation in dynamic environments and lack mechanisms to retain and reuse previously acquired knowledge. While Continual RL (CRL) seeks to develop systems capable of lifelong learning and adaptation, existing methods and evaluation benchmarks are severely constrained by prohibitive computational costs and unrealistic assumptions. Furthermore, training agents to exhibit complex behaviors still relies heavily on carefully handcrafted tasks.
This thesis addresses these critical limitations to advance the development of autonomous agents capable of continual, real-world lifelong learning. Specifically, the main objectives are to: **(I)** design scalable CRL architectures that retain and reuse knowledge throughout an agent's lifetime under minimal assumptions; **(II)** establish evaluation frameworks that accurately replicate the real-world challenges faced by lifelong learning agents; and **(III)** develop a framework for the autonomous generation of increasingly complex behaviors without human intervention.

Keywords: Reinforcement Learning · Continual Learning · Open-Endednes.

1 Introduction

This document presents the doctoral thesis proposal of Mikel Malagón, which started on the 1st of February 2024 at the University of the Basque Country (UPV/EHU) under the supervision of Dr. Jon Vadillo, Dr. Josu Ceberio, and Prof. Jose A. Lozano.

The field of Reinforcement Learning (RL) focuses on developing models—referred to as agents—that learn directly from experience [28]. These agents interact with an environment by taking actions that alter its state, thereby influencing the data available for subsequent learning. Typically, agents are trained through trial and error to maximize a cumulative scalar reward over time, which quantifies the quality of the selected actions relative to a predefined task. This perception-action-reward loop between the agent and the environment is illustrated in Figure 1. At each discrete timestep, the agent observes the current

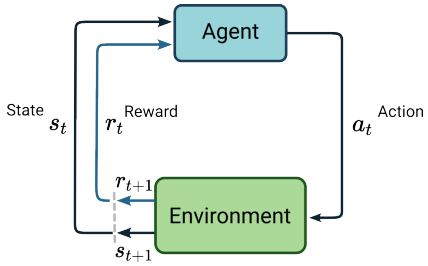


Fig. 1. Diagram of the agent-environment interaction loop commonly considered in RL.

environment's state and selects an action, resulting in a state transition and a corresponding reward. Over time, the agent aims to discover an optimal behavior policy that maximizes the expected long-term cumulative reward.

Over the past decade, the use of deep Neural Networks (NNs) in RL has led to numerous breakthroughs not only within the RL field but also across various scientific and engineering domains [21,2,34]. For instance, in [26], deep RL was deployed to discover an accelerated sorting algorithm, while in [17], an RL-based system predicted protein folding structures with unprecedented accuracy.

However, most of these powerful deep RL algorithms are typically trained from scratch (starting with random parameters) to solve a specific, well-defined task under a stationarity assumption—i.e., that the environment will never change after training [16,20]. While many research-oriented environments satisfy this criterion, this is rarely the case in real-world scenarios, where the actual deployment of most of these models is intended to occur. For instance, an agent trained to perform a mechanical task—such as controlling a robotic arm or a production line mechanism—implicitly assumes that the physical properties of the machine are invariant. This is an unrealistic assumption given that mechanical parts inevitably degrade over time. Another example involves controlling a real-world water treatment plant [15]; because the water source is natural, its short- and long-term properties depend on highly unpredictable factors like weather conditions or natural events, making the stationarity assumption invalid. Moreover, when faced with changing environmental dynamics or shifting goals, conventional RL algorithms not only fail to adapt but can fail catastrophically, entirely losing their capacity to learn [7]. In contrast, humans and other animals seamlessly adapt to changing environments, greatly benefiting from past experiences to efficiently tackle novel scenarios that arise over time. In this realm, Continual Reinforcement Learning (CRL) aims to develop agents capable of continuous adaptation and endless learning [1].

However, many RL methods—especially those based on deep NNs—suffer dramatic performance degradation due to well-known issues such as catastrophic forgetting [7], interference [18], and loss of plasticity [22]. *Catastrophic forgetting* [10] occurs as learning a new task with an NN requires modifying its parameters; as the network adjusts to the new objective, it can override parameters crucial for previously learned tasks, causing performance on earlier tasks to degrade

dramatically. *Interference* happens when the parameters learned in one task make optimizing a subsequent task considerably more difficult [18]. Starting a new task with parameters optimized for a previous one introduces an extremely biased initialization, which can result in lower performance than initializing the network randomly (the standard practice in conventional RL scenarios). Finally, *loss of plasticity* occurs when the parameters of the NN are updated in such a way that the model permanently loses its capacity for further learning and adaptation [22].

Over recent years, the field of CRL has introduced various methods to address these core limitations [18]. However, many existing approaches suffer from the stability-plasticity dilemma—the trade-off between maintaining the capacity to learn and retaining knowledge [25]—or lack the mechanism to effectively reuse prior experiences [7]. One prominent class of methods designed to mitigate these challenges is growable NN architectures [27,14]. These approaches prevent catastrophic forgetting and the loss of plasticity by dynamically instantiating new NN modules as required, enabling effective knowledge transfer via shared representations. Nevertheless, they require prohibitively high computational and memory costs.

Consequently, the main objective of the initial phase of this thesis was to reduce the computational costs associated with these methods. To this end, we introduced CompoNet, a novel growing NN architecture built upon self-composing policy modules [23]. We showed that CompoNet effectively leverages previously acquired knowledge to solve novel tasks while preventing interference between them and maintaining plasticity across a suite of robotic control tasks [33] and Atari environments [3]. This work was subsequently published and selected for an oral presentation at the International Conference on Machine Learning (ICML) 2024.

Beyond CRL methods, another prominent challenge in the field is the lack of suitable environments for designing and evaluating continually learning agents. For instance, conventional CRL benchmarks predominantly focus on the task-based view of CRL, requiring an agent to maximize cumulative rewards within a sequence of distinct tasks that change abruptly over time [33,30]. This piecewise-stationary perspective is highly restrictive and fails to capture the complexity of most real-world lifelong learning scenarios [8]. In practice, the environment's dynamics may evolve continuously, or nonstationarity may naturally arise from the perspective of a computationally limited agent, an idea known as the Big World hypothesis [16,29]. Although the literature features popular Big World-like benchmarks such as MineRL [13] (based on Minecraft) and NetHack [19], both present significant limitations. MineRL suffers from a prohibitive computational cost and relies on the closed-source commercial game Minecraft, while NetHack is a text-based ASCII game that completely overlooks the challenges of 3D worlds vital for physical real-world interactions, such as autonomous driving or robotic manipulation. Furthermore, most RL environments, including NetHack and MineRL, offer minimal support for modifying the environment's properties beyond basic settings or building entirely custom scenarios. We argue that the capac-

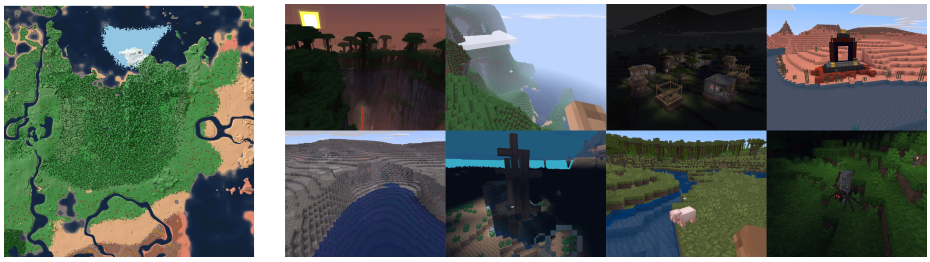


Fig. 2. The rightmost figures showcase of some the predefined Craftium environments, while the leftmost plot illustrates a top-view of the open-world Craftium environment.

ity to design custom environments is a crucial feature, as it allows researchers to systematically verify specific properties of the agents in detail, such as 3D navigation capabilities, memory, or robustness.

To address these limitations, the second work presented in this thesis introduces Craftium [24], an open-source platform designed to balance flexibility and efficiency for rich 3D RL environments, see Figure 2. Published at the International Conference on Machine Learning (ICML) 2025, Craftium is built on a custom fork of the Luanti voxel engine [3] that natively integrates the popular Gymnasium and PettingZoo RL APIs [31] to support both single- and multi-agent setups. By eliminating the heavy Java runtime overhead typical of Minecraft-based alternatives, it achieves a substantial improvement in performance of over 2K steps per second as shown in Figure 3. The platform has rapidly gained community traction, achieving almost 200 GitHub stars[4] at the time of this writing and earning adoption in top-tier literature (e.g., ICLR and ICML 2026) and usage in prominent institutions like Microsoft Research [6,11] and Google DeepMind [5].

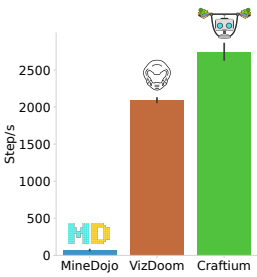


Fig. 3. Performance comparison between MineDojo (Minecraft-based), VizDoom (not 3D), and Craftium in a single-environment single-agent setting (higher is better).

While the first two contributions focus on methods and evaluation frameworks for CRL, a fundamental question remains: how can agents autonomously develop complex behaviors without relying on manually engineered tasks? To address this question, the third work of this thesis shifts the focus toward a foundational study of autonomous learning. Currently under review, this work formalizes a new notion of efficient exploration and analyzes its theoretical and empirical implications. Specifically, we demonstrate that this definition of exploration is sufficient on its own to drive an open-ended process that naturally gives rise to an automatic curriculum of increasingly complex

[3] https://www.luanti.org/en/.
[4] See https://github.com/mikelma/craftium.

behaviors in the complete absence of any reward function or external objectives and tasks.

Finally, current and future work is directed toward two key lines of research: (1) optimizing the computational and memory efficiency of the CompoNet architecture while further minimizing its underlying assumptions, and (2) establishing a complementary benchmark environment that isolates the challenges of CRL. Regarding the first effort, while CompoNet significantly improves upon prior approaches, we believe that its scalability can be improved by integrating state-of-the-art policy distillation techniques and leveraging recent insights into mitigating the loss of plasticity in deep NNs. Regarding the second, although Craftium provides a rich and flexible infrastructure for complex environments, it introduces orthogonal challenges such as sparse rewards and high-dimensional representation learning. Developing a complementary Big World environment that strictly isolates the challenges of lifelong learning will accelerate foundational algorithmic research, which can later be combined with specialized methods to tackle the full environment complexity of Craftium.

2 Objectives

The main goal of this thesis is to address the critical limitations that hinder the deployment of continually learning agents in real-world environments. For the sake of guiding the project towards the aforementioned goal, in the subsequent lines, we outline the three specific sub-objectives.

O1 Proposing a scalable growing NN architecture. Regarding Objective (I), while existing growing NNs can mitigate catastrophic forgetting, they suffer from a quadratic parameter cost relative to the number of network modules, hindering their deployment in real-world scenarios [27]. Existing mitigation strategies attempt to lower these costs by limiting expansion or packing new information into existing parameters; however, this introduces a restrictive trade-off between minimizing computational overhead and maintaining the capacity to acquire new knowledge (i.e., plasticity) [14,12]. To resolve this dilemma, we developed *CompoNet*, a modular architecture that leverages advanced attention mechanisms [32] and representation learning to dynamically compose and reuse previously learned modules [23]. By routing information through relevant sub-networks, CompoNet can leverage forward knowledge transfer while achieving a linear parameter growth rate with respect to the number of modules. This represents a significant scaling improvement over the quadratic costs of prior approaches, enabling agents to maintain plasticity and reuse past knowledge over long lifetimes. This work was published at the *Forty-first International Conference on Machine Learning (ICML)* and was selected for oral presentation (top 5% of accepted papers).

O2 Developing tools for analyzing CRL algorithms. Existing CRL evaluation benchmarks predominantly consist of sequences of tasks; the agent's

objective is to maximize performance on the current task, and after some predefined number of timesteps, the task changes abruptly [33,30]. This view of CRL is very limited and does not represent the challenges of continual learning in real life, where nonstationarities are mostly smooth, and the agent cannot always exploit hard task transitions [8]. Moreover, the environments that most accurately reproduce the challenges of real-world CRL—e.g., MineRL [13] and MineDojo [9]—imply prohibitive computational cost and rely on the closed-source Minecraft game and do not allow extensive modifications to existing environments or creating custom ones. Consequently, we developed Craftium, an open-source framework for creating 3D voxel-based environments for RL agents [24].[5] Craftium allows creating extremely broad and diverse environments (such as in Minecraft) but improves the efficiency of the environments by orders of magnitude (see Figure 3). Moreover, Craftium allows extending, modifying, or creating new environments at an unprecedented granularity thanks to its flexible Lua API, an extended version of the Luanti API [6]. Finally, Craftium was published at the *Forty-second International Conference on Machine Learning (ICML 2025)* and was also presented in the *Reinforcement Learning and Video Games (RLVG) Workshop* at the *Conference of Reinforcement Learning 2025*, where it was awarded the best paper award.

O3 Foundations for the autonomous generation of complex behaviors. While the first two contributions focus on methods and evaluation frameworks for CRL, a fundamental question remains: how can agents autonomously develop complex behaviors without relying on manually engineered tasks? To address this question, the third work of this thesis shifts the focus toward a foundational study of autonomous learning. Currently under review, this work formalizes a new notion of efficient exploration and analyzes its theoretical and empirical implications. Specifically, we demonstrate that this purely intrinsic objective is sufficient to drive the emergence of highly sophisticated behaviors by framing exploration around prediction and model generalization. Rather than maximizing uniform environment coverage, we define efficient exploration as prioritizing the generation of generalizable experience that minimizes the global prediction error of a learned world model over the long run. Theoretically, we show that under standard Markovian assumptions, optimally efficient explorers naturally schedule their trajectories to visit the most informative and learnable regions first, thereby avoiding absorbing states and the classic noisy-TV problem [4]. Empirically, we utilize a Monte Carlo approximation method to optimize agents and world models parameterized by NNs in partially observable settings. Our results demonstrate that this optimization process naturally induces an automatic curriculum of progressively more complex behaviors—such as exploiting environmental symmetries, solving mazes, and employing advanced navigation

[5] Craftium's main page: https://craftium.readthedocs.io/.
[6] See https://api.luanti.org/.

strategies. Ultimately, this framework offers a principled mechanism for sustaining an open-ended process of autonomous learning without relying on external rewards, tasks, or manually engineered objectives.

3 Methodology and Work Plan

This section describes the general methodology that will be followed during the thesis project. We review the objectives presented in Section 2 under the described methodology, specifying sub-tasks accordingly. The general methodology proposed for the achievement of each of the objectives consists of five phases, described in what follows.

1. **Statement of the problem:** This first phase consists of identifying, characterizing, and mathematically formalizing the problem.
2. **Analysis of the state of the art:** Consists of searching, reading, and documenting publications on RL and CRL.
3. **Method design and experimentation:** Finding a methodological solution to the problem, as generic as possible, so that it can later be applied in different settings. A general solution to the problem is proposed, and suitable experimentation is designed and executed to evaluate and validate this proposal. This phase requires the usage of the computational resources of the research group, such as cluster nodes of several GPUs.
4. **Application of the methodology to real problems:** The general methodology designed in the previous stage is applied to one or several problems extracted from environments of real interest, and the results are analyzed. In this phase, it may be necessary to make small adaptations to the generic methodology, to fit the application's scope.
5. **Publication of results:** Submission of scientific papers to prestigious journals and congresses in the area. Normally, these correspond to journals with impact factor, those belonging to the first quartile (according to the JCR), or to international conferences in the field of machine learning qualified as A++ (according to the SCIE ranking). In addition, the submission of papers with preliminary results to national congresses in the area (of lower rating) is valued.

To illustrate the methodology above for each of the objectives of the project, the following is a description of each of the phases of the project.

O1 Proposing a scalable growing neural network architecture.

T1.1: Mathematical formalization of the CRL setting.

T1.2: Exhaustive literature review on growing NN architectures and catastrophic forgetting mitigation strategies.

T1.3: Implementation of a modular architecture utilizing advanced attention mechanisms for parameter-efficient module composition (CompoNet [23]).

T1.4: In-depth architectural analysis and ablation studies to evaluate individual module contributions to overall performance and parameter growth.

T1.5: Design of a comprehensive experimental validation pipeline across robotic control and high-dimensional tasks.

T1.6: Large-scale empirical benchmarking against state-of-the-art CRL methods to validate our approach.

T1.7: Publication and presentation of the proposed method in a top-tier ML venue (work published at ICML 2024, and selected for oral presentation, top 5% of the accepted papers).

O2 Developing tools for analyzing CRL algorithms.

T2.1: Characterization of the overheads and limitations in existing CRL evaluation frameworks.

T2.2: State-of-the-art review of RL environments and simulators, procedural generation techniques, and standardized interfaces.

T2.3: Design and implementation of an open-source framework for creating rich 3D voxel-based environments (Craftium [24]).

T2.4: Analyze execution speed and throughput profiling to validate the proposal.

T2.5: Open-sourcing the codebase, including ready-to-use environments and publishing the work top ML venue (work published at ICML 2025).

O3 Foundations for the autonomous generation of complex behaviors.

T3.1: Mathematical formalization of the problem and the definition of efficient exploration.

T3.2: Exhaustive literature review on intrinsic motivation, world models, and reward-free exploration.

T3.3: Theoretical analysis of the presented mathematical framework (under reasonable assumptions, such as Markovianity).

T3.4: Design of a practical optimization pipeline utilizing a Monte Carlo approximation method.

T3.5: Empirical analysis of the complex emergent behaviors induced by the efficient exploration.

T3.6: Writing a technical paper and submission to a top-tier ML conference (currently under review at NeurIPS 2026).

Finally, since the development of the project does not involve any experimentation or direct interaction with humans or living creatures, we do not see any major ethical implications for this work. However, as the impact of AI systems in the real world continues to grow, we believe in an open and ethical development of such systems. Therefore, as a common task of all three objectives, we will develop a public free and open-source repository with all code, data, and instructions necessary to replicate the experiments. In this way, the experiments conducted will be transparent, comparable, and reproducible for future work.

4 Relevance

This thesis establishes a comprehensive paradigm for advancing lifelong learning agents by bridging algorithmic innovation, evaluation infrastructure, and foundational exploration theory. To date, the first two primary objectives have been successfully completed and validated at top-tier ML venues. Specifically, Objective (I) resulted in the publication of CompoNet at ICML 2024 (selected for oral presentation, top 5% of the accepted papers), showing that modular NNs architectures can achieve linear parameter growth while retaining plasticity and long-term knowledge transfer capabilities. Objective (II) led to the development of Craftium, an open-source, high-throughput 3D platform published at ICML 2025, which eliminates severe infrastructure bottlenecks for evaluating RL agents. Expanding upon these structural frameworks, Objective (III) addresses the foundational challenge of the autonomous generation of complex behaviors through an intrinsic exploration framework currently under review at NeurIPS. Together, these contributions offer a unified, principled approach to sustaining open-ended continual learning without relying on manually engineered tasks or external rewards.

Beyond its scientific contributions, the unified framework of this thesis carries significant potential for real-world societal application and technology transfer. Traditional RL systems remain notoriously brittle when deployed in dynamic environments, suffering from devastating performance degradation or catastrophic forgetting when physical conditions change or operational conditions drift. By providing a scalable architecture capable of continuous adaptation, a high-fidelity platform for rigorous pre-deployment stress testing, and a mechanism for reward-free autonomous skill acquisition, this research facilitates the deployment of resilient autonomous systems in the real world. These capabilities are crucial for next-generation industrial automation, long-duration robotic operations, and adaptive decision-making networks, where environmental fluctuations are inevitable, and hand-crafting external reward functions is economically or logistically prohibitive.

Looking forward, the final phase of this thesis builds upon a three-month research visit (from March 1st to May 31st, 2026) at the University of Alberta, the center leading global RL research. Working alongside Prof. Michael Bowling and the students from the Reinforcement Learning and Artificial Intelligence (RLAI) group, directed by Turing Award winner Rich Sutton, we are currently pursuing two collaborative initiatives to further develop these contributions. First, we are improving CompoNet's computational efficiency by incorporating policy distillation techniques and selective forgetting. Second, we are establishing a complementary benchmark within the "Big World" paradigm to run alongside Craftium. While Craftium provides the rich infrastructure needed for full environmental complexity, isolating orthogonal challenges of CRL like sparse rewards and representation learning in a targeted environment will accelerate core algorithmic refinement.

References

1. Abel, D., Barreto, A., Van Roy, B., Precup, D., van Hasselt, H.P., Singh, S.: A definition of continual reinforcement learning. Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) **36**, 50377–50407 (2023)
2. Bellemare, M.G., Candido, S., Castro, P.S., Gong, J., Machado, M.C., Moitra, S., Ponda, S.S., Wang, Z.: Autonomous navigation of stratospheric balloons using reinforcement learning. Nature **588**(7836), 77–82 (2020)
3. Bellemare, M.G., Naddaf, Y., Veness, J., Bowling, M.: The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research (JAIR) **47**, 253–279 (2013)
4. Burda, Y., Edwards, H., Storkey, A., Klimov, O.: Exploration by random network distillation. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019)
5. Castanyer, R.C., Mohamed, F., Castro, P.S., Neary, C., Berseth, G.: ARM-FM: Automated reward machines via foundation models for compositional reinforcement learning. arXiv preprint arXiv:2510.14176 (2025)
6. Cohen, L., Nabati, O., Wang, K., Kumar, N., Mannor, S.: Horizon imagination: Efficient on-policy rollout in diffusion world models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2026)
7. Dohare, S., Hernandez-Garcia, J.F., Lan, Q., Rahman, P., Mahmood, A.R., Sutton, R.S.: Loss of plasticity in deep continual learning. Nature **632**(8026), 768–774 (2024)
8. Elelimy, E., Szepesvari, D., White, M., Bowling, M.: Rethinking the foundations for continual reinforcement learning. In: Reinforcement Learning Journal (RLJ) (2025)
9. Fan, L., Wang, G., Jiang, Y., Mandlekar, A., Yang, Y., Zhu, H., Tang, A., Huang, D.A., Zhu, Y., Anandkumar, A.: MineDojo: Building open-ended embodied agents with internet-scale knowledge. Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) **35**, 18343–18362 (2022)
10. French, R.M.: Catastrophic forgetting in connectionist networks. Trends in cognitive sciences **3**(4), 128–135 (1999)
11. Garcin, S., Walker, T., McDonagh, S., Pearce, T., Bilen, H., He, T., Wang, K., Bian, J.: Beyond pixel histories: World models with persistent 3d state. In: Proceedings of the International Conference on Machine Learning (ICML) (2026)
12. Gaya, J.B., Doan, T., Caccia, L., Soulier, L., Denoyer, L., Raileanu, R.: Building a subspace of policies for scalable continual learning. In: Proceedings of the International Conference of Learning Representations (ICLR) (2023)
13. Guss, W.H., Houghton, B., Topin, N., Wang, P., Codel, C., Veloso, M., Salakhutdinov, R.: MineRL: a large-scale dataset of minecraft demonstrations. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). pp. 2442–2448 (2019)
14. Hung, C.Y., Tu, C.H., Wu, C.E., Chen, C.H., Chan, Y.M., Chen, C.S.: Compacting, picking and growing for unforgetting continual learning. Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) **32** (2019)
15. Janjua, M.K., Shah, H., White, M., Miahi, E., Machado, M.C., White, A.: Gvfs in the real world: making predictions online for water treatment. Machine Learning **113**(8), 5151–5181 (2024)
16. Javed, K., Sutton, R.S.: The big world hypothesis and its ramifications for artificial intelligence. In: Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks (2024)

17. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al.: Highly accurate protein structure prediction with alphafold. Nature **596**(7873), 583–589 (2021)
18. Khetarpal, K., Riemer, M., Rish, I., Precup, D.: Towards continual reinforcement learning: A review and perspectives. Journal of Artificial Intelligence Research (JAIR) **75**, 1401–1476 (2022)
19. Küttler, H., Nardelli, N., Miller, A., Raileanu, R., Selvatici, M., Grefenstette, E., Rocktäschel, T.: The NetHack learning environment. Proceendings of the Advances in Neural Information Processing Systems (NeurIPS) **33** (2020)
20. Lewandowski, A., Ramesh, A., Meyer, E., Schuurmans, D., C Machado, M.: The world is bigger! a computationally-embedded perspective on the big world hypothesis. Proceedingd of the Advances in Neural Information Processing Systems (NeurIPS) **38**, 28210–28237 (2026)
21. Li, S.E.: Reinforcement learning for sequential decision and optimal control. Springer (2023)
22. Lyle, C., Zheng, Z., Nikishin, E., Pires, B.A., Pascanu, R., Dabney, W.: Understanding plasticity in neural networks. In: Proceedings of the International Conference on Machine Learning (ICML) (2023)
23. Malagon, M., Ceberio, J., Lozano, J.A.: Self-composing policies for scalable continual reinforcement learning. In: Proceedings of the International Conference on Machine Learning (ICML) (2024)
24. Malagón, M., Ceberio, J., Lozano, J.A.: Craftium: Bridging flexibility and efficiency for rich 3d single-and multi-agent environments. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 42765–42794 (2025)
25. Mallya, A., Lazebnik, S.: Packnet: Adding multiple tasks to a single network by iterative pruning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7765–7773 (2018)
26. Mankowitz, D.J., Michi, A., Zhernov, A., Gelmi, M., Selvi, M., Paduraru, C., Leurent, E., Iqbal, S., Lespiau, J.B., Ahern, A., et al.: Faster sorting algorithms discovered using deep reinforcement learning. Nature **618**(7964), 257–263 (2023)
27. Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R.: Progressive neural networks. arXiv preprint arXiv:1606.04671 (2016)
28. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press (2018)
29. Sutton, R.S., Koop, A., Silver, D.: On the role of tracking in stationary environments. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 871–878 (2007)
30. Tomilin, T., Fang, M., Zhang, Y., Pechenizkiy, M.: COOM: A game benchmark for continual reinforcement learning. Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) **36**, 67794–67832 (2023)
31. Towers, M., Kwiatkowski, A., Balis, J., De Cola, G., Deleu, T., Goulão, M., Andreas, K., Krimmel, M., Kg, A., Perez-Vicente, R., et al.: Gymnasium: A standard interface for reinforcement learning environments. Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) **38** (2026)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) (2017)
33. Wołczyk, M., Zając, M., Pascanu, R., Kuciński, Ł., Miłoś, P.: Continual world: A robotic benchmark for continual reinforcement learning. Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) **34** (2021)

34. Wurman, P.R., Barrett, S., Kawamoto, K., MacGlashan, J., Subramanian, K., Walsh, T.J., Capobianco, R., Devlic, A., Eckert, F., Fuchs, F., et al.: Outracing champion gran turismo drivers with deep reinforcement learning. Nature **602**(7896), 223–228 (2022)